DOCUMENT RESUME

ED 247 294 TM 840 475

AUTHOR Beard, Jacob G.; And Others
TITLE Domain-Referenced Tests: Achieving Equivalence Through Rasch Pre-Equating.
PUB DATE Apr 84
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Criterion Referenced Tests; *Difficulty Level; *Equated Scores; Latent Trait Theory; Secondary Education; Test Items
IDENTIFIERS *Domain Referenced Tests; Domain Specifications; *Rasch Model; Test Equivalence

ABSTRACT

The purpose of this study was to examine the homogeneity in difficulty of item domains and the effectiveness of Rasch pre-equating procedures for adjusting test scores for differences in the difficulty of tests constructed by sampling from item domains. The data used were taken from a field test and calibration of 810 tenth-grade items in mathematics and communications skills. Approximately 250 students responded to each of the items. The mean range of item difficulty for 15 mathematics skills domains was 2.593 logits and for 10 communications skills domains was 4.210 logits. It was found that Rasch pre-equating techniques effectively adjusted total scores from tests of 45 and 53 items, selected from specific sub-domains, for substantial differences in the level of difficulty of items included in the tests. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED247294

Domain-Referenced Tests: Achieving Equivalence Through Rasch Pre-Equating

Jacob G. Beard
Gail Fletcher
Lucinda Richards

Florida State University

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

[X] This document has been reproduced as received from the person or organization originating it.

[ ] Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Fletcher

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of AERA, New Orleans, 1984.

PRECIS

DOMAIN-REFERENCED TESTS: ACHIEVING EQUIVALENCE THROUGH RASCH PRE-EQUATING

Jacob G. Beard, Gail Fletcher and Lucinda Richards

The purpose of this study was to examine the homogeneity in difficulty of item domains and the effectiveness of Rasch pre-equating procedures for adjusting test scores for differences in the difficulty of tests constructed by sampling from item domains. The mean range of item difficulty for 15 mathematics skills domains was 2.593 logits and for 10 communications skills domains was 4.210 logits. It was found that Rasch pre-equating techniques effectively adjusted total scores from tests of 45 and 53 items, selected from specific sub-domains, for substantial differences in the level of difficulty of items included in the tests.

Domain-Referenced Tests: Achieving Equivalence Through Rasch Pre-Equating

Jacob G. Beard, Gail Fletcher &
Lucinda Richards
Florida State University

INTRODUCTION

The purpose of this study was to examine the homogeneity in difficulty of item domains and the effectiveness of Rasch pre-equating procedures for adjusting test scores for differences in the difficulty of tests constructed by sampling from item domains.

With the advent of criterion-referenced testing, increasingly greater emphasis has come to be placed on domain definition for the generation of test items. Domain-referenced testing is based upon the assumption that by identifying the domain through the specific definition of a class of behaviors, skills and information, a set of rules may be created for generating sets of test items. It has been argued that where the domain has been sufficiently well defined, a random sample of items can provide a means of achieving an absolute measure of domain proficiency (Kriewall, 1972). Given that the pool of acceptable items has been carefully delineated, then as L. S. Smith (1978) points out, it follows that "those items which are actually generated should realize a high degree of `definitional homogeneity'" (pg. 24). Definitional homogeneity is promoted by item generation schemes which involve some logical, systematic and replicable means for constructing items representative of the defined domain. Whether response patterns to such items display homogeneity, however, is one of the main topic of this investigation.

Williams and Slawski (1980) concluded that two tests developed from the same set of specifications and presumably meeting Popham's (1978) criterion of equivalence (randomly equivalent) produced different score distributions and would lead to different interpretations of student performance. Their analyses were done using scores from short competency tests of five and ten items each. On the other hand, a study reported by the Northwest Evaluation Association (1978), in which longer tests were used and the scores were equated using Rasch techniques led to the conclusion that "... students do receive comparable achievement levels from different tests, provided that [none of the tests] is at a grossly inappropriate level."

How much variability can be expected among the items specified for a domain with definitional homogeneity? This is a

very important consideration because student performance on such test items is sometimes used to make far-reaching decisions about the student. These decisions should not be dependent upon which particular set of items from the domain were included in the test.

Hence, the purpose of this paper was to investigate the following questions:

1. How variable in difficulty are items generated from item specifications for singularly defined domains?

2. How parallel are alternative forms of tests which are constructed by sampling items of varying difficulties from several defined domains?

3. Can differences in the difficulty of alternative forms of domain-referenced tests be compensated for by pre-equating the scores using the Rasch model?

METHODOLOGY

Instruments

The instruments used in the study were generated from two sources. In order to answer the question regarding the variability in difficulty of items yielded by sets of item specifications, a subset of the items from an item pool was examined. This part of the item pool was the product of an item writing effort in which 810 items were written for 25 domains, each defined by a single set of item specifications conforming generally to the form described by Popham (1978). For questions 2 and 3, the pre-existing item responses from an eighth-grade test of minimum competency in basic skills were used. The item responses were from a random sample of eighth-grade students who took the test during its statewide administration in 1981. The test contained 152 communications items and 114 mathematics items. However, for the purposes of this study, shorter tests were generated by selecting particular subsamples of items from the full-length tests.

Subjects

The data used in describing the similarity of items from a domain were taken from a field test and calibration of 810 tenth-grade test items. The 810 items were divided into 25 forms of approximately 32 items each. The 25 forms were spiraled and administered in 16 schools in 8 districts selected to represent the achievement level and variance of the state. These forms were administered along with the statewide tests which served as an anchor test for calibrating the items to the common statewide scale of difficulty. Approximately 250 students responded to each of the items.

For questions 2 and 3, the data were taken from a random sample of the eighth-grade students who took the statewide basic skills test in the fall of 1981.

Procedures

For the first question, descriptive statistics were used to show the mean, standard deviation, and range of the items' difficulties within each of the 25 domains.

To address questions 2 and 3, the following procedures were used. From the complete eighth-grade mathematics test of 114 items, five subtests of 45 items each were selected using the following sampling procedure. For the first test, the two most difficult items of each of the 25 domains (skills) were chosen. The second test was formed by selecting the second and third most difficult items; the third test by selecting the third and fourth most difficult items; the fourth test by selecting the fourth and fifth most difficult items (which for most skills were the two easiest items). A fifth test was formed by randomly selecting two items from each skill area. For the one skill having other than five items, the sampling plan was adjusted to yield items from the analogous portion of the item difficulty continuum. This process was repeated for the communications test, which yielded five subtests of 53 items each.

After the items for the five tests in each of the two subject matter areas had been selected, the student item response records were rescored to yield "number right" scores for each student on each of the ten tests. This procedure yielded scores showing the number of items which each student would have answered correctly if he had taken the ten subtests.

The means, standard deviations, and intercorrelations among the raw scores of the five tests within each subject area were computed. Additionally, the raw scores of each test were pre-equated to a common scale and the same statistics computed.

The term "pre-equating" as used here means that the Rasch log difficulties computed for the intact tests of 114 and 152 items were used to calculate the raw to log score transformations for each of the shorter structured tests. The item responses from the full tests were used to compute actual raw scores for each subject for each structured test. This is different from the usual use of the term "pre-equating" in which non-overlapping sets of subjects are used to derive item difficulties.

These procedures yielded, for each student, five raw scores for mathematics and five for communications; five equated scaled scores for mathematics and five for communications. It was assumed that the raw scores would vary by an amount that could have been predicted from the mean difficulties of the items included. It was hypothesized that the mean equated scaled scores would not be significantly different because the equating

process should have adjusted the scores for the differences among the difficulties of the five forms.

However, the similarity of scaled score means would not in itself constitute an argument that the tests were parallel in function. In order to address the question of parallelism the product-moment correlations among the different forms of the tests were compared with the square root of the product of their reliabilities, the maximum value for the expected correlation in each case.

RESULTS

Variability of Items Within Domains

The variability of item difficulties generated from the same and from different sets of item specifications are shown in Table 1 and in Figures 1 and 2. It is obvious from the data that items generated from the same set of item specifications were not equivalent nor even similar in difficulty. The mean range of difficulty, in logits, adjusted to a common scale within the Mathematics and within Communications, was 2.55 logits for Mathematics and 4.25 for Communications. It is also apparent that, for many skills, there is almost as much variance of difficulties within skills as between skills.

The Mathematics skill which yielded items having the greatest range in difficulty was: "Subtract two mixed numbers with denominators of 2,3,4,5,6,8,or 10." The two most extreme items were examined and the differences in difficulty were easily explained. The whole number portion of both was two digits. The easiest item required no conversion of the whole number or fractional portions for subtraction, while the most difficult item required regrouping for both the whole number and fractional subtraction and the calculation of a least common denominator. It is not surprising that the odds of a correct response are many times greater for the easy item than for the difficult one.

The Communications skill yielding items having the greatest range in difficulty was: "Spell words needed in writing through grade 10". Again, the two items differing most in difficulty were examined. The easiest item required the selection of the correctly spelled word "fortunate" from four alternatives including three obviously incorrect alternatives. The most difficult item required the selection of the correctly spelled word "maintenance" from four choices including three subtle variations. The variations in spelling included two which were based on the form of the root word maintain. Rasch methodology indicates that the odds of spelling the easier of these two words in 170 times greater than for the more difficult one.

These data show the substantial differences among the difficulties of items presumably assessing the same skill. In order to achieve homogeneity of item difficulties it would be necessary to construct hundreds of narrowly defined objectives

leading to an ultra-operationalism in reporting of scores, which would be extremely difficult to use with parents and the lay public.

Adjusting for Varying Difficulties of Domain-Referenced Tests Through Pre-Equating

When tests were constructed by sampling items of different difficulties, the raw score means varied as predicted, but the pre-equated scaled score means were identical. The mean raw scores, shown in Table 2, reflect the item sampling procedure used in constructing the different tests. The mean raw scores are smallest for the test composed of the most difficult items, and increase systematically to the test composed of the easiest items. Test 5, constructed by randomly sampling items from the domains yielded raw score means between the two extremes in difficulty and closer to those of the more difficult tests.

The pre-equating techniques were effective in adjusting for the differences in difficulty of the items. When the scaled score means were rounded to the nearest whole number scaled score, they were identical for the mathematics and for the communications tests. This degree of similarity was unexpected in that Test 1 was composed of the two hardest items in each domain and Test 4 was composed of the two easiest items of each domain. These two tests differed in difficulty by almost a standard deviation on the raw score scale in each case, mathematics and communications. These results suggest that equating to an equated score scale would accurately adjust scores for differences in level of difficulty arising from the selection of a particular set of items from defined domains.

The reliabilities (alpha coefficients) are shown for the raw scores of each test as the diagonal values of the correlation matrix in Tables 3 and 4. These values are similar for the different tests with a tendency for the most difficult test to be slightly less reliable.

In addition to similarity in mean scale scores, it is necessary that the alternate forms of the test produce scores which are highly correlated in order for them to be considered parallel. The correlations among the different forms are shown in Tables 3 and 4. Several generalizations can be made from these correlations.

First, the correlations between the raw and scaled scores of each form range from .93 to .97, reflecting the small degree of non-linearity in the transformation. It is apparent that raw scores generally correlate higher with the raw scores of other test forms than with their scaled scores, and that scaled scores correlate higher with scaled scores than with raw scores.

Second, the correlations between Test 5 scores, structured

by randomly sampling items, and other test scores are, with few exceptions, as high as their reliabilities permit. Because Test 5 tended to be closer to the difficult end of the series of tests they behaved more like those tests rather than the easier tests.

Third, for comparisons between any two specific tests raw scores generally yielded slightly higher correlations than did scaled scores.

Fourth, scaled scores from tests which were only one step apart in difficulty were generally correlated as highly as their reliabilities would permit for mathematics (.94, .93 and .94) and only slightly less for communications (.90, .90, .89). Scores from tests which were two steps apart in difficulty yielded correlations of .85 and .87 for mathematics and .85 and .84 for communications. When the tests were three steps apart in difficulty the correlations were .83, for mathematics and .82 for communications. It should be remembered that the latter values occurred when the two easiest items from a domain were selected for one test and the two hardest items from the domain were selected for the other.

SUMMARY AND CONCLUSIONS

The purpose of this study has been to contribute to our knowledge of item domains and to examine some procedures for working with them. It was found that items, written to the same specifications, vary considerably in difficulty. The degree of heterogeneity in difficulty is so great that single items or small item samples from a domain can not be considered parallel in any useful sense. This was especially true for domains of items in the subject area of communications skills, and only slightly less true for domains in the area of mathematics.

It was found that Rasch pre-equating techniques effectively adjusted mean scores from tests of 45 and 53 items, for substantial differences in the difficulty of items included. The equated scaled score means of the tests were found to be identical, even when the raw score means differed by nearly one standard deviation. The correlations of scaled scores were substantial in all cases and large enough to consider the tests as parallel when the raw score means differed by no more than one-third of a standard deviation.

The use of the pre-equating or equating techniques could solve a major problem faced by large-scale minimum competency testing programs. It is apparent that randomly parallel tests having a small number of items may not be equivalent. Attempts to control the difficulty level of the tests through purposive sampling of items on the basis of difficulty, may lead to restricted coverage of content. The data of this study support the notion that items could be selected to maximize content validity, even if their difficulty values were moderately

different and the tests could be effectively pre-equated to yield similar results. On the other hand, if domains of content are broadly defined, one must seriously question the parallelism of tests differing widely in difficulty.

Table 1

Mean, Standard Deviation, and Range of Item Difficulties for Each of Twenty-five Item Specifications

Mathematics				Communications			
Skill	Mean	S.D.	Range	Skill	Mean	S.D.	Range
D32	-.72	.37	1.45	I70	-.74	.70	2.73
C24	-.69	.97	3.57	D20	-.56	1.02	4.16
G56	-.36	.46	1.64	D17	-.37	.75	3.55
K78	-.32	.68	2.81	I75	-.34	.93	4.11
I69	-.05	.52	2.01	A2	-.31	.99	4.11
B17	.13	.50	1.14	G61	.20	1.34	5.14
I70	.27	.96	2.74	G58	.20	.99	4.35
L80	.28	.39	1.72	C14	.35	1.04	4.98
B16	.37	.97	4.34	B8	.67	1.10	5.05
H62	.45	.49	2.01	H69	1.74	.87	4.28
I71	.53	1.01	4.47				
L81	.63	.46	2.14				
D34	.77	.68	2.68				
K79	.82	.99	3.75				
I72	1.92	.40	1.72				
Mean range			2.55	Mean range			4.25

TABLE 2

MEANS, STANDARD DEVIATIONS AND RELIABILITIES OF SCORES FROM THE MATHEMATICS AND COMMUNICATIONS TESTS

Test	Mean[1]	Standard Deviation	Reliability
MATHEMATICS[2]			
1 Raw	32.69	7.56	.91
1 Scaled	413	27.04	
2 Raw	35.05	7.16	.92
2 Scaled	413	28.54	
3 Raw	36.74	6.65	.92
3 Scaled	413	29.06	
Raw	38.00	6.07	.92
Scaled	413	30.10	
5 Raw	35.83	6.80	.92
5 Scaled	413	29.39	
COMMUNICATIONS[3]			
1 Raw	40.21	8.03	.91
1 Scaled	415	26.01	
2 Raw	43.52	7.86	.92
2 Scaled	415	28.59	
3 Raw	45.51	7.33	.93
3 Scaled	415	29.52	
4 Raw	47.12	7.16	.95
4 Scaled	415	31.18	
5 Raw	42.77	7.63	.92
5 Scaled	415	28.31	

[1] n = 801 for mathematics and 692 for communications
[2] k = 45
[3] k = 53

TABLE 3

INTERCORRELATIONS AMONG THE RAW AND SCALED SCORES
FOR THE FIVE MATHEMATICS TESTS

	PEARSON CORRELATION COEFFICIENTS									
	(Hardest) TEST 1		TEST 2		TEST 3		(Easiest) TEST 4		(Random) TEST 5	
	Raw	Scaled	Raw	Scaled	Raw	Scaled	Raw	Scaled	Raw	Scaled
1R	.91*	.97	.96	.94	.89	.87	.87	.85	.93	.92
1S			.91	.94	.83	.85	.81	.83	.89	.92
2R			.92	.96	.95	.92	.91	.88	.95	.91
2S					.90	.93	.85	.87	.90	.92
3R					.92	.96	.96	.92	.95	.89
3S							.90	.94	.91	.91
4R							.92	.95	.94	.88
4S									.91	.91
5R									.92	.96
5S										

* alpha reliabilities on the diagonal

TABLE 4

INTERCORRELATIONS AMONG THE RAW AND SCALED SCORES
FOR THE FIVE COMMUNICATION TESTS

	PEARSON CORRELATION COEFFICIENTS									
	(Hardest) TEST 1		TEST 2		TEST 3		(Easiest) TEST 4		(Random) TEST 5	
	Raw	Scaled	Raw	Scaled	Raw	Scaled	Raw	Scaled	Raw	Scaled
1R	.91*	.97	.93	.90	.88	.87	.84	.86	.95	.93
1S			.87	.90	.81	.85	.75	.82	.90	.93
2R			.92	.95	.95	.90	.91	.88	.94	.89
2S					.87	.90	.80	.84	.90	.90
3R					.93	.94	.95	.91	.93	.86
3S							.84	.89	.90	.89
4R							.95	.93	.89	.79
4S									.89	.85
5R									.92	.96
5S										

* alpha reliabilities on the diagonal

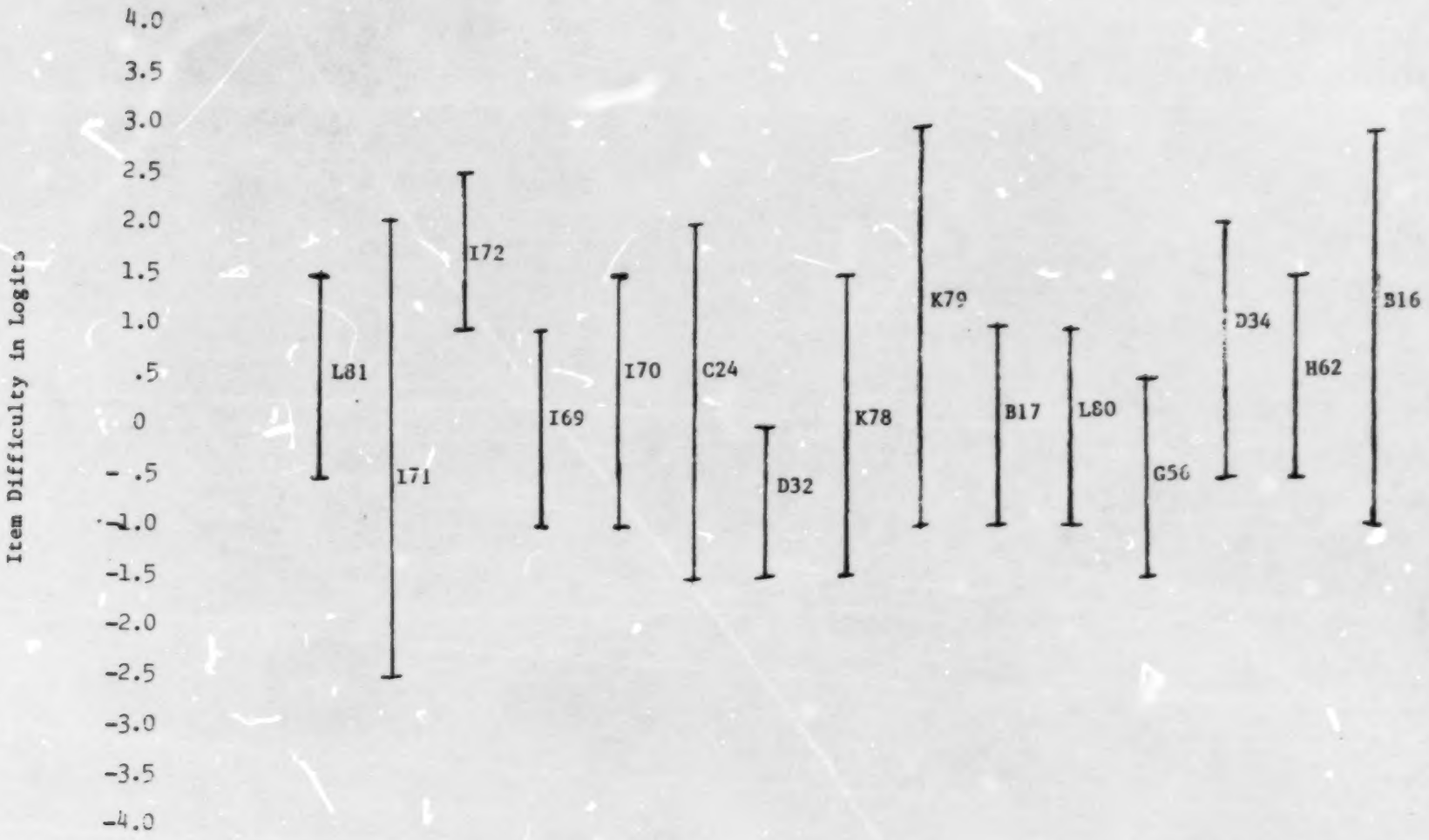

Figure 1. Range of Difficulty for each Skill: Mathematics.

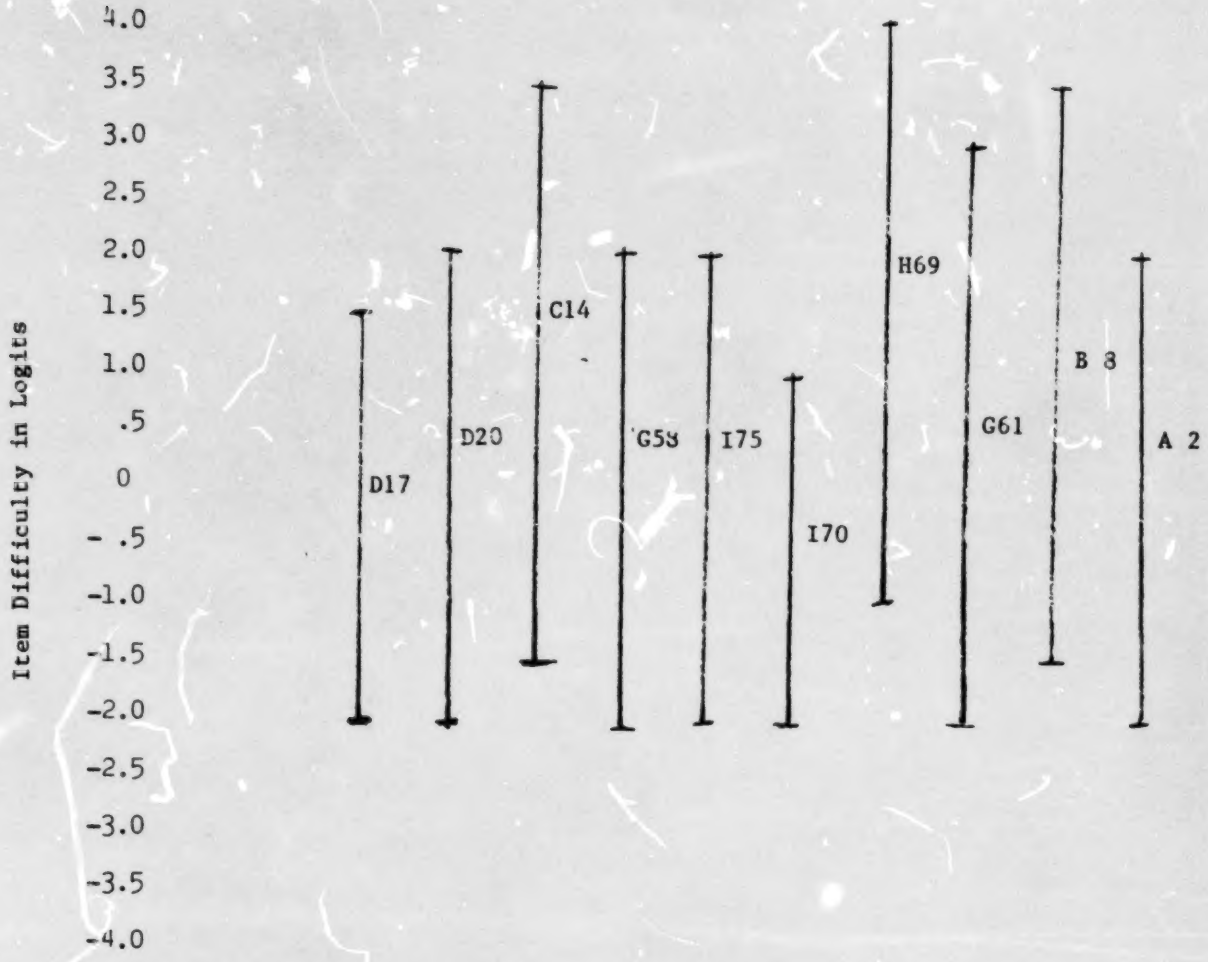

Figure 2. Range of Difficulty for each Skill: Communications.

REFERENCES

Bigler, E., & Williamson, M. (Eds.). (1978). Do students receive the same Rasch achievement level from different tests? [Rasch Model Monograph Series]. Northwest Evaluation Association Journal, Monograph IV, 13-16.

Kriewall, T E. (1972). Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, February 1972, Chicago.

Popham, W.J. (1978). Criterion referenced measurement. Englewood, N.J.: Prentice-Hall.

Smith, L.S. (1978). Decisions and dilemmas in constructing criterion-referenced test: Some questions and issues. (CSE Report No. 110). Washington, D.C.: National Institute of Education.

Williams, P.L., & Slawski, E.J. (1980). Applications of the Rasch model for the development of equivalent forms of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, 1980, Boston.